Automatisierte Erstellung von Erkrankungsmodellen mit gesundheitsökonomischer Verwendung am Beispiel eines Tumorregisters - Erste Voruntersuchungen

Dipl.-Inform. Med. Monika Pobiruchin

GECKO Institut für Informatik, Medizin und Ökonomie
Hochschule Heilbronn
Max-Planck-Str. 39
74081 Heilbronn
monika.pobiruchin@hs-heilbronn.de

Abstract: Gesundheitsökonomische Modelle gehören mittlerweile zum Standardwerkzeug in der Beurteilung der Wirtschaftlichkeit von neuen Therapien. Sie stehen dabei in einem besonderen Zwiespalt, zum einen sollen sie komplexe Sachverhalte darstellen, aber zum anderen auch transparent und nachvollziehbar für die Entscheidungsträger sein. Erkrankungsmodelle entstehen heute noch vorwiegend manuell, die nötigen Strukturen und Wahrscheinlichkeiten werden der Literatur oder klinischen Studien entnommen. Gerade hier liegt jedoch die Schwierigkeit: Ergebnisse von klinischen Studien können nur bedingt auf die Alltagsroutine übertragen werden.
Diese Arbeit soll mit Hilfe von Algorithmen aus dem Bereich des Data Minings die Erstellung von Modellen beschleunigen und erleichtern. Das fertige Modell soll auf Basis von realen Daten aus einem klinischen Krebsregister erstellt werden, statt sich alleine auf publizierten Daten aus der Literatur zu stützen.

1 Einleitung

Erkrankungsmodelle werden in der Gesundheitsökonomie als Standard-Analysetechnik für Wirtschaftlichkeitsprüfungen eingesetzt. Dabei wird das Verhältnis zwischen dem medizinischen Nutzen und Kosten einer bestimmten Therapie gegenüber einem Vergleichsstandard untersucht. Als Datenquellen werden für die Modellierung vor allem klinische Studien, Registerstudien und systematische Übersichtsarbeiten genutzt. Viele dieser Datenquellen müssen mangels Verfügbarkeit im deutschsprachigen Raum aus dem internationalen Umfeld entlehnt werden. Sie können nur bedingt auf das deutsche Gesundheitswesen übertragen werden.

Daten aus der klinischen Routine, Bsp. klinische Patientenakten, werden in der Modellierung nur selten berücksichtigt. Ein Grund hierfür ist, dass erst in jüngster Vergangenheit an den medizinischen Einrichtungen Strukturen geschaffen worden sind, die die Daten aus verschiedenen Subsystemen der Krankenhaus-IT zusammenführen und für wissenschaftliche Fragestellungen bereithalten. Ein weiterer Grund besteht in der Notwendigkeit von

Langzeitdaten für die Modellierung, die nicht in allen Einrichtungen in bereits elektronischer und auswertbarer Form vorliegen.

Die gesundheitsökonomische Betrachtung von Krebserkrankungen bzw. deren Therapien wird in den folgenden Jahren immer mehr an Bedeutung gewinnen. Gerade in Hinblick auf die entstehenden Kosten, die im Jahr 2008 in Deutschland 18 Milliarden Euro betrugen. Dies waren 7% der gesamten Krankheitskosten [Bun10]. Bereits heute gibt es zahlreiche Evaluationen auf dem Gebiet des Brustkrebs, laut [KH05] jedoch nicht für Deutschland. Grundlage für die Evaluationen bilden gute, transparente und nachvollziehbare Modelle. Ziel dieses Forschungsprojekts ist es dies mit Methoden des Data Minings zu erreichen.

2 Hintergrund

2.1 Gesundheitsökonomische Modelle

In dem speziellen Feld der Gesundheitsökonomie ist ein Modell nach der Definition der ISPOR (International Society for Pharmacoeconomics and Outcomes Research) eine Methode, die Ereignisse im Zeitverlauf in Bezug einer Population darstellt und zur Entscheidungsunterstützung dient. Modelle sollen die Effekte von bestimmten Interventionen in Bezug auf die Gesundheit der Population und die Auswirkungen auf die Kosten sichtbar machen [WOH$^+$03].

Nach [Ake03] ist die Modellierung immer dann notwendig, wenn über eine Erkrankung lediglich Daten über einen kurzen Zeithorizont vorliegen und eine informationelle Lücke in Bezug auf einen langfristigen Ausblick klafft. Nicht immer sind die Auswirkungen und Einflüsse von neuen Maßnahmen und Therapien den Entscheidungsträgern klar ersichtlich. Modelle können hier einen wertvollen Beitrag liefern und Zusammenhänge aufzeigen. Dabei greifen sie nicht in die reale Welt ein. Veränderungen an Modellparametern bedingen unmittelbar eine Veränderung der Ergebnisse. Ein Vorgehen, das unter realen Umständen so nicht möglich ist.

Eine in der Gesundheitsökonomie häufig genutzte Methode zur Darstellung von Modellen ist nach [BCS06] die Markov-Modellierung. Ein Markov-Modell besteht aus einer Anzahl n Zuständen $q_1, q_2, ...q_n$, die miteinander durch Übergänge verbunden sind. Dabei sind rechnerisch $n * n$ - Übergänge möglich, die in einer Übergangsmatrix dargestellt werden. Anhand der Struktur von Markov-Modellen lassen sich sehr gut pathologische Prozesse ableiten und verfolgen.

2.2 Mining von Assoziationen und Sequenzen

Das Auffinden von Assoziationen in vorliegenden Datenmengen gehört nach [Liu11] seit Anfang der 90ziger Jahre zu einem der Hauptforschungsgebiete des Data Minings. Ein Beispiel aus dem Bereich der Krebserkrankungen soll hier statt des oft bemühten Warenkorbs für die Erläuterungen herangezogen werden. Eine Assoziationsregel kann ausgedrückt werden mit $PositiverTastbefund- > Biopsie[Support = 10\%, Konfidenz = 80\%]$

Dies bedeutet, das bei 10% der gesamten Fälle ein positiver Tastbefund und eine Biopsie zusammen auftreten. Bei 80% der Patientinnen folgt auf einen positiven Tastbefund noch eine Biopsie der Brust. Auch wenn es bei diesem Beispiel den Anschein erwecken mag, dass die zeitliche Abfolge berücksichtigt wird, bei dem assoziierendem Mining wird keinerlei Unterscheidung nach Zeitpunkt gemacht. $PositiverTastbefund- > Biopsie, Laboruntersuchung$ wäre gleichbedeutend mit
$PositiverTastbefund- > Laboruntersuchung, Biopsie.$

Bei dem sogenannten Sequential Pattern Mining wird zusätzlich noch die zeitliche Abfolge der Items berücksichtigt. Eine Sequenz ist dabei eine geordnete Liste von Items. Die Items werden einer Datenbasis, dem Itemset, entnommen. Wenn I ein Itemset mit $I = \{a, b, c, d, e, f, g, h\}$ ist, dann ist beispielsweise $s =< \{a\}\{c, d\}\{b\}\{a, d\} >$ eine Sequenz. $\{c, d\}$ kann auch als Transaktion bezeichnet werden.

Der GSP-Algorithmus (Generalized Sequential Patterns) findet häufige Subsequenzen in dem er die Datenbasis zuerst nach einzelnen häufig auftretenden Items durchsucht. Ausgehend von den einelementigen Sequenzen, werden zweielementige Sequenzen gebaut (Join step) und ihre Häufigkeit in Bezug auf das gesamte Itemset festgestellt. Tritt eine Sequenz nicht häufig genug auf, d.h. ihr Support ist nicht hoch genug, wird sie wieder verworfen (Prune step). Aus den verbleibenden zweielementigen Sequenzen wird wiederum versucht Sequenzen mit drei Elementen zu bilden, etc..

3 Methoden

Das in Kapitel 2 skizzierte Sequenzmining soll in den nächsten Arbeitsschritten auf eine relevante Kohorte, die aus den Datensätzen des Tumorregisters gezogen werden, angewandt werden. Dabei soll gezeigt werden, dass Patienten mit den selben diagnostizierten Erkrankungen auch einen vergleichbaren klinischen Verlauf aufweisen. Die gefundenen Sequenzen sind die Grundlage für die Struktur des gesundheitsökonomischen Modells. Ob die Sequenzen und Strukturen mit dem tatsächlichen klinischen und pathologischen Geschehen korrespondieren, wird ein Vergleich mit gängigen Leitlinien oder den im Klinikum implementierten Behandlungspfade zeigen.

Die Wahrscheinlichkeit der Übergänge zwischen den einzelnen Zuständen des Markov-Modells soll ebenfalls aus den extrahierten Sequenzen geschätzt werden. Hierbei muss nicht ausschließlich ein Maximum Likelihood Schätzer zum Einsatz kommen wie in der Machbarkeitsstudie (s. Abschnitt 3.1). Andere Herangehensweisen zum Schätzen der besten Parameterwerte wie z.B. der Maximum Entropy Approach wären denkbar. Auch hier können bereits entwickelte Modelle wie in [LHJ+09] herangezogen werden, um mögliche Abweichungen oder Übereinstimmungen aufzudecken.

3.1 Machbarkeitsstudie

Um dieses Vorgehen zu untermauern, wurde eine Machbarkeitsstudie mit einem anonymisierten Abrechnungsdatensatz vorgenommen, der den Spezifikationen nach §21 des Krankenhausentgeltgesetzes (Krankenhausentgeltgesetz vom 23. April 2002 (BGBl. I S. 1412, 1422), das zuletzt durch Artikel 7 des Gesetzes vom 22. Dezember 2011 (BGBl. I S. 2983) geändert worden ist.) entspricht[1]. Der Datensatz umfasste 9.726 Patienten und 16.099 Prozeduren. Er bildete einen Zeitraum von 13 Monaten ab (November 2005 bis Dezember 2006).

Die §21-Datensätze lagen als CSV-Dateien vor. In eine relationale Datenbank importiert, konnten mittels SQL-Abfragen relevante Kohorten gezogen werden.

Für die Voruntersuchung wurde die Open Source - Software Rapid Miner [MWK+06] eingesetzt. Neben Rapid Miner wurde auch das Softwaretool Weka (Waikato Environment for Knowledge Analysis [HFH+09]) getestet, doch die Anwendung von Weka erwies sich als mühsam. Besonders die Einschränkung, dass Weka lediglich das ARFF-Dateiformat (Attribute Relationship File Format) unterstützt, wurde als Hemmnis empfunden. Die Algorithmen des Wekapakets können jedoch als Erweiterungen in Rapid Miner integriert werden, so dass bei der Benutzung von letzterem keinerlei Einbußen in Bezug auf die zur Verfügung stehenden Analysealgorithmen hingenommen werden müssen.

Die Kohorte bestand aus Frauen mit Brustkrebs (ICD-Code C50) als Haupt- oder Nebendiagnose. Im vorliegenden Zeitraum des §21-Datensatzes wurden die vorgenommenen und abgerechneten Prozeduren als Itemset I betrachtet. Aus diesem Itemset wurden mit Hilfe des GSP-Algorithmus Sequenzen gebildet. Bei der Erstellung der Modellstruktur wurden zwei einschränkende Annahmen getroffen: Die Übergangswahrscheinlichkeiten sind zeitlich unabhängig. Der Übergang in einen bestimmten Zustand ist lediglich abhängig vom zuvor gewählten Zustand (Markov-Modell der 1. Ordnung). Die Wahrscheinlichkeiten für den Übergang in einen Zustand ist somit bestimmt durch $P(q_{i+1}|q_i)$. Ein Zustand im Markov-Modell wurde hier definiert als eine Transaktion in der Sequenz. Mittels eines Maximum Likelihood Schätzers wurden die Übergangswahrscheinlichkeiten bestimmt.

[1] Im folgenden als §21-Datensatz bezeichnet.

4 Erste Ergebnisse

Die betrachtete Kohorte umfasste 96 Patientinnen an denen 41 verschiedene Prozeduren vorgenommen wurden. Für eine Patientin ergab sich so z.B. die Sequenz $s_1 =< \{5-87\} >< \{5-89\} >$. Dies bedeutet, dass an einem Tag eine „Exzision und Resektion der Mamma“ an einem späteren, darauffolgenden Tag eine „Operation an Haut und Unterhaut“ bei dieser Patientin durchgeführt wurde. Bei einem Minimumsupport von 20% ergaben sich fünf verschiedene Sequenzen, die insgesamt 137-mal auftraten. Die resultierende Modellstruktur zeigt Abb. 1.

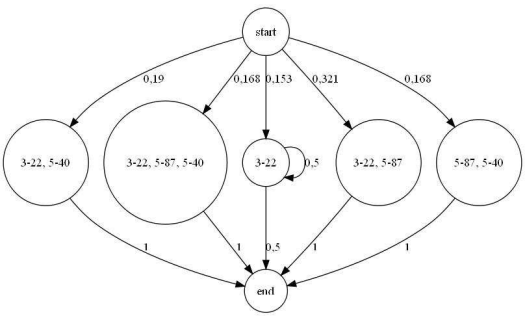

Abbildung 1: Modellstruktur mit Übergangswahrscheinlichkeiten. $3-22$: Computertomographie (CT) mit Kontrastmittel, $5-40$: Operation am Lymphgewebe, $5-87$: Exzision und Resektion der Mamma, $5-89$: Operation an Haut und Unterhaut.

In fast allen Zuständen wird zunächst eine radiologische Untersuchung mittels Computertomographie vorgenommen, dann folgen diverse Eingriffe. Anhand der Zustände lässt sich erkennen, dass diese jeweils an einem Tag vorgenommen wurden. CT-Aufnahmen werden beispielsweise bei einem Verdacht auf Lungenmetastasen empfohlen. Das Modell zeigt auch, dass manche Frauen gar nicht operiert wurden (ca. 15%).

5 Diskussion und Ausblick

Einschränkend muss jedoch angemerkt werden, dass es sich bei den vorliegenden Daten um keine Längsschnittdaten handelt. Die gefundenen Sequenzen waren mit zwei bis drei Elementen entsprechend kurz. Der §21-Datensatz umfasste lediglich Abrechnungsdaten in einem Zeitraum von etwas über einem Jahr. Eine Krankengeschichte lässt sich daher nur bedingt mit diesen Datensätzen abbilden. Es können auch keine Aussagen zu zeitlichen Zusammenhängen gemacht werden, die Übergangswahrscheinlichkeiten des Modells in Abb. 1 stehen in keinem zeitlichen Kontext, sondern leiten sich lediglich von den absoluten Zahlen ab.

Der §21-Datensatz wies keine Daten zu Chemotherapien bei Brustkrebs auf. Insgesamt

muss geurteilt werden, dass dieser vorliegende Datensatz sicher nicht geeignet ist ein Modell der Brustkrebserkrankung darzustellen. Doch hier sei noch einmal darauf hingewiesen, dass dies auch nicht das Ziel der Machbarkeitsstudie war.

Die ca 40.000 Datensätze im Tumorregister, die für die kommenden Arbeitsschritte verwendet werden, reichen hingegen bis in die 80er Jahre zurück und beinhalten neben einer Basisdokumentation auch Einträge zu verschiedensten Behandlungen und Befunde. Dies sollte eine fundierte Grundlage sein, um Verläufe der Patienten abbilden zu können.

Es gibt bereits aktuelle Forschung im Schnittfeld von Krebserkrankungen (insbesondere auch Brustkrebs) und Data Mining. Diese beschäftigen sich jedoch häufig mit der Vorhersage von Krebserkrankungen wie [AdMC12] oder dem Erfolg einer bestimmten Therapie wie [TSO+12] und bewegen sich weniger im Kontext der Gesundheitsökonomie. In Taiwan war es bei [SPT09] möglich auf Basis des nationalen Versicherungsprogramms und der dort gespeicherten Abrechnungsdaten eine gesundheitsökonomische Analyse zur Kosteneffektivität von bestimmten Chemotherapieschemata durchzuführen. Die Ergebnisse dieser Untersuchung standen im Widerspruch zu ähnlichen Evaluationen. [SPT09] führte dies auf den Umstand zurück, dass sie ihre Berechnungen auf reale Abrechnungsdaten statt auf Studienergebnisse stützten. Jedoch wurde von den Autoren für die Untersuchung kein Markov-Modell gewählt.

Die zukünftigen Arbeiten dieses Forschungsprojekts sollen diese Lücken schließen und ein Markov-Modell als Ergebnis vorweisen, dem reale Behandlungsdaten als Grundlage dienen.

Literatur

[AdMC12] A. Aussem, S. R. de Morais und M. Corbex. Analysis of nasopharyngeal carcinoma risk factors with Bayesian networks. *Artif Intell Med*, 54(1):53–62, Jan 2012.

[Ake03] R. L. Akehurst. Making decisions on technology availability in the British National Health Service–why we need reliable models. *Value Health*, 6(1):3–5, 2003.

[BCS06] A. Briggs, K. Claxton und M. J. Sculpher. *Decision modelling for health economic evaluation*. Oxford University Press, 2006.

[Bun10] Statistisches Bundesamt. Krankheitskosten - 2002, 2004, 2006 und 2008. *Fachserie 12 Reihe 7.2*, 2010.

[HFH+09] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann und I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.

[KH05] R. Kath und M. Hartmann. Gesundheitsökonomische Evaluation des Mammakarzinoms. *Der Onkologe*, 11:152–163, 2005.

[LHJ+09] M. P. Lux, M. Hartmann, C. Jackisch, G. Raab, A. Schneeweiss, K. Possinger, J. Oyee und N. Harbeck. Cost-utility analysis for advanced breast cancer therapy in Germany:

results of the fulvestrant sequencing model. *Breast Cancer Res Treat*, 117(2):305–17, 2009.

[Liu11] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Springer; 2nd Edition, 2011.

[MWK+06] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz und T. Euler. YALE: Rapid Prototyping for Complex Data Mining Tasks. In Lyle Ungar, Mark Craven, Dimitrios Gunopulos und Tina Eliassi-Rad, Hrsg., *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Seiten 935–940, New York, NY, USA, August 2006. ACM.

[SPT09] Y. T. Shih, I. P. und Y. Tsai. Information technology facilitates cost-effectiveness analysis in developing countries: an observational study of breast cancer chemotherapy in Taiwan. *Pharmacoeconomics*, 27(11):947–961, 2009.

[TSO+12] M. Takada, M. Sugimoto, S. Ohno, K. Kuroi, N. Sato, H. Bando, N. Masuda, H. Iwata, M. Kondo, H. Sasano, L. W C Chow, T. Inamoto, Y. Naito, M. Tomita und M. Toi. Predictions of the pathological response to neoadjuvant chemotherapy in patients with primary breast cancer using a data mining technique. *Breast Cancer Res Treat*, Jun 2012.

[WOH+03] M. C. Weinstein, B. O'Brien, J. Hornberger, J. Jackson, M. Johannesson, C. McCabe, B. R. Luce und I. S. P. O. R. Task Force on Good Research Practices-Modeling Studies. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices–Modeling Studies. *Value Health*, 6(1):9–17, 2003.